Evaluation of software to map NGS reads from heterogeneous HCV1b populations

Lize Cuypers[1], Joke Snoeck[1,*], Bram Vrancken[1], Lien Kerremans[1,#], Grégoire Vuagniaux[2], Frederik Nevens[3], Anne-Mieke Vandamme[1,4]

[1]Clinical and Epidemiological Virology, Rega Institute for Medical Research, KU Leuven, Leuven, Belgium [2]Debiopharm SA, Lausanne, Switzerland [3]Department of Hepatology, University Hospitals Leuven, Belgium [4]Centro de Malária e Outras Doencas Tropicais and Undidade de Microbiologia, Universidade Nova de Lisboa, Lisbon, Portugal * currently at Nuffield Department of Surgical Sciences, John Radcliffe Hospital, Oxford, United Kingdom # now at Multiplicom N.V., Niel, Belgium

Background

Direct-acting antivirals (DAAs) for HCV treatment achieve high success rates, but are more likely to select for viral drug-resistant mutants. DAA combination therapies targeting different viral proteins are used to tackle this problem. In order to detect resistance mutations simultaneously in all viral target genes under selective pressure, a full-genome assay was developed. Both population and next-generation sequencing were performed, but the analysis of NGS data from diverse viral populations remains a non-trivial exercise. In order to accommodate the specific characteristics of short read fragment data, the impact of tailored analysis methodologies and of reference sequence divergence on the recovery of minority variants, was studied.

Material and methods

Four samples from treatment-naïve HCV1b infected patients were amplified by nested PCR and sequenced with Illumina's Genome Analyzer IIx. The performance of four software packages for aligning reads against three reference sequences was tested:

- a published HCV1b reference sequence (AB049087)
- a consensus of the respective sample obtained by Sanger sequencing
- an *in silico* reconstructed data-specific reference sequence (VICUNA)

Concordance between the viral population after Sanger sequencing and after read mapping with Segminator II, was compared. An approximate maximum-likelihood tree was constructed to evaluate the covered variability spectrum.

Results

Performance of packages to align reads to a reference sequence

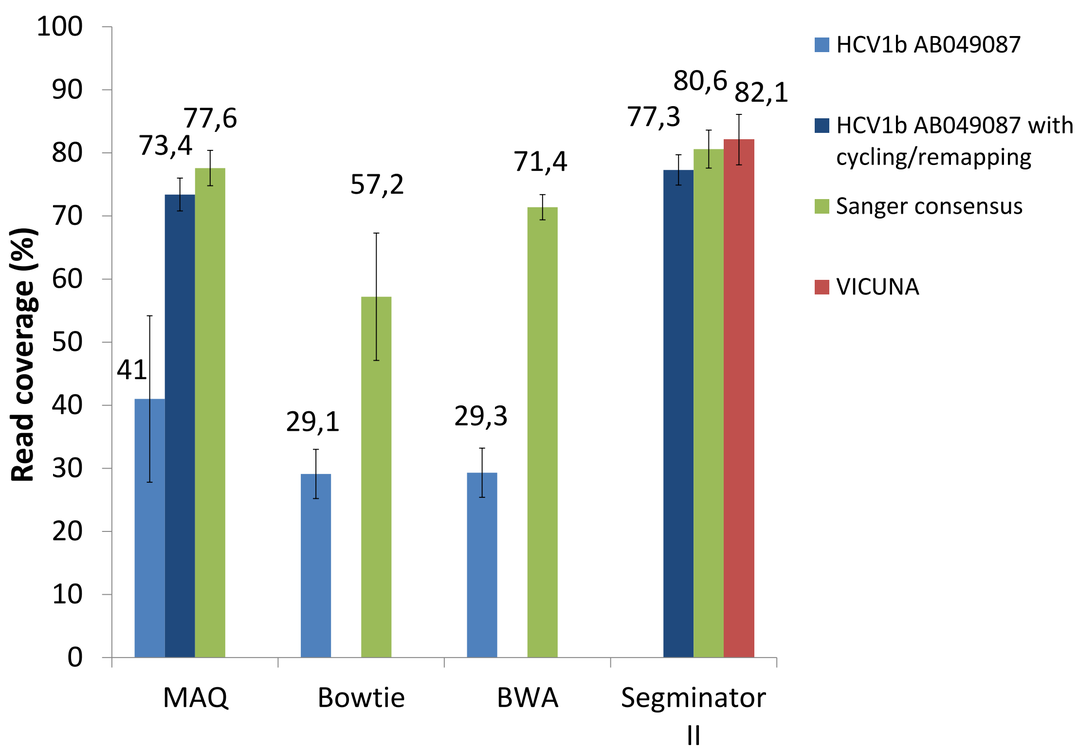

Figure 1. Mapped reads. Software packages (MAQ, Bowtie, BWA and Segminator II) were used to align reads to reference sequences: HCV1b AB049087 (in blue), Sanger consensus (in green) and *in silico* data-specific reference (only Segminator, in red). For MAQ the AB049087 reference was used to build a contig which served as reference in the next rounds of mapping (cycling, in dark blue) and in case of Segminator II, the template was automatically replaced with the obtained consensus during mapping (remapping).

For all packages, the number of mapped reads was consistently higher when a sample-specific sequence was used instead of the HCV1b AB049087 reference. Segminator II was the best software package in recovering number of reads: 80,6% (± 3.0) against the sample-specific Sanger sequence, 77.3% (± 2.4) against the AB049807 reference and 82.1% (± 4.0) against an *in silico* data-specific reference.

Impact of the reference sequence divergence

Reference sequence	Major + minor discordance (%, stdev)	Major discordance (%, stdev)
HCV1bAB049087 (Genbank)	1,72 (+/- 0,791)	0,30 (+/- 0,092)
Sanger sequence (sample-specific)	1,70 (+/- 0,700)	0,026 (+/- 0,014)
Reconstructed *in silico* reference (data-specific)	1,47 (+/- 0,790)	0,05 (+/- 0,024)

Table 1. Discordances between Sanger and NGS sequencing. Comparison between the NGS consensus sequences obtained by Segminator II (remapping), using all three reference sequences, and the sample-specific sequence obtained with Sanger sequencing. Minor discordances are predominantly attributable to ambiguity characters in the NGS consensus. Major discordances are differences in nucleotides or are due to indels of one nucleotide.

Simple scoring of aligned positions as (un)matched revealed a high degree of concordance between NGS consensus and sample-specific Sanger sequences. After neglecting minor discordances, overall concordance increased to a nearly perfect match (>99%). Closer inspection of the differences in minority variants revealed that with NGS they were mainly present below the 0.5% threshold. However, differences at frequencies above 5% were all located at (70.34%) or within 5 nucleotides of a divergent area, defined as areas with Shannon entropy above the 75th percentile.

Spectrum of variability covered by the assay

The wide dispersion of the HCV1b clade over the deepest branches illustrates that this assay captures most of the variability.

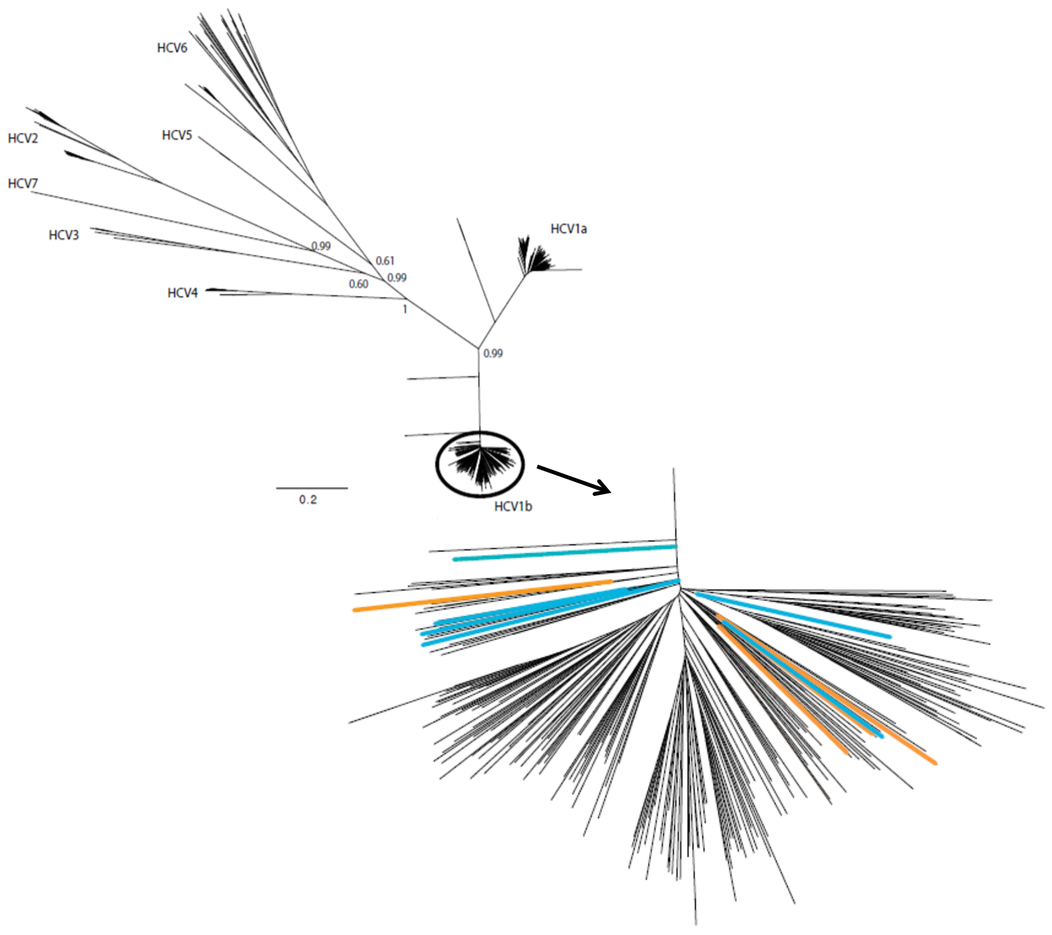

Figure 2. Maximum-likelihood tree with FastTree, 1000 bootstraps. Ten samples sequenced by Sanger sequencing (blue) and four Illumina sequences (orange), together with a reference alignment of LANL including all HCV genotypes. The respective genotype clades are indicated.

Conclusions

Only a few software packages can accommodate the diversity present in heterogeneous viral populations. Using a distantly related reference sequence leads to substantial data loss, even with the best among the packages, Segminator II. Although the consensus sequence is largely independent of the used reference sequence, when discordances (mostly differences in mixtures) do occur, they are enriched in highly divergent regions. So the use of sample-specific reference sequences for read mapping, is recommended.

Acknowledgments

This work was supported by a postdoctoral grant (PDOH/09/FWO) and a PhD grant (Asp/12, 1102513N) of the FWO. Part of this research was sponsored by an OT grant (OT/08/047) from the Research Fund KULeuven, by an FWO grant (G.A029.11), an FWO account (1.5.252.12N) and by Debiopharm SA. This research has also received funding from the European Union Seventh Framework Programma [FP7/2007-2013] under Grant Agreement nr 278433-PREDEMICS and ERC Grant Agreement nr 260864.